MODELING WORD DURATION FOR BETTER SPEECH RECOGNITION

Venkata Ramana Rao Gadde

Speech Technology and Research Lab
SRI International, Menlo Park, CA 94025, USA
Email: rao@speech.sri.com

ABSTRACT

We describe a new method of modeling duration at word level. These duration models are easily trained from the acoustic training data and can be used to rescore N−best lists of recognition hypotheses. The models capture some of the well known durational effects such as prepausal lengthening. They incorporate a simple back off mechanism to handle unseen words during rescoring. Experiments with various large vocabulary conversational speech recognition (LVCSR) evaluation sets showed consistent improvements of 0.7−1.0% in word error rate (WER).

1. INTRODUCTION

Current research in speech recognition is focussed on segmental features such as Mel frequency cepstra (MFC), which perform well in noise−free environments. However, their performance degrades significantly in the presence of noise. They are also susceptible to channel variations. Prosodic features − duration, energy and pitch − are more robust to noise and are unaffected by channel conditions. Hence, it is important to develop prosodic features and models to improve speech recognition. Past research in duration (or, in general, prosody) modeling focused on the use of prosodic features to aid in syntactic and semantic analysis of speech. Examples of these studies are [1] and [2]. Another study [3] reported that word models incorporating lexical stress perform better than models without stress. Some recent studies reported the use of prosody to detect hidden events in conversational speech and showed that it results in improved speech recognition [4]. However, none of these tried to develop explicit models of prosody that can be automatically trained from a training set and can be used for speech recognition along with segmental models. Our study addresses these issues in that the word duration models we propose are easily trained from the acoustic training data and can be easily integrated into the standard hidden markov model (HMM) based recognizers. A similar approach was reported in a recent paper on duration modeling [5].

2. WORD DURATION MODELS

The basic idea of the word duration models was described at the Hub5 1999 workshop [6]. Each word is represented by a duration feature that is a vector comprising of the durations of the individual phones in the word. For example, the word "that", represented as the phone sequence "dh+ae+t", may be represented by a duration feature (10.0 8.0 4.0), where the three values, 10.0, 8.0 and 4.0 represent the durations of the three phones, "dh", "ae", and "t" respectively. Thus, the duration feature captures the durations of the phones within the context of the given word. Given sufficient instances of a word, that is, sufficient number of duration features, we can train statistical models to represent the word duration. In our experiments, we found it convenient to use Gaussian mixture models (GMMs), although other models are possible. These word duration GMMs can then be used to rescore the recognition hypotheses in an N−best list.

In developing the word duration models, we addressed three issues. First, we considered the issue of unseen words. Since the word duration models were trained from the acoustic training data, the models are limited to the words in the training vocabulary. In addition, we train a model for a word only if there are a minimum number of occurrences of the word

in the training data. Hence, it is possible to encounter words for which we do not have a model during recognition. To handle this, we train duration models of individual triphones and phones along with those of words. We implemented a simple back off scheme, in which the triphone models are used to rescore the unseen word. If a triphone model does not exist, we back off to the corresponding phone model.

The second issue was the durational effect known as 'prepausal lengthening' which refers to the lengthening of the syllables of a word preceding a pause. To incorporate this into the models, we modified our training to train separate models for words followed by another word, and words followed by a pause.

The third issue relates to normalization of rate of speech (ROS) across different speakers. We computed ROS as the average number of phones per second over an utterance. The ROS was then used to normalize the durations of the phones. We tried two ways of using the ROS normalization, at the hypothesis level and at the speaker level, and compared the performance with the unnormalized performance. Hypothesis−level normalization was found to give inconsistent results across different test sets, whereas speaker−level normalization gave a consistent win on all test sets.

3. EXPERIMENTS

We first performed forced alignment of all the 220,000 utterances in the LVCSR acoustic training data against their transcriptions. From these alignments, we obtained the durations of the phones for all the words in the training data. Using these, we trained the word duration models. We also trained the triphone and phone duration models for back off purposes.

For testing we used two sets, the 1996 LVCSR male eval set (EVAL96) and our 2000 development set (DEV00), which was a subset of the Hub5 1998 eval set. The EVAL96 set was used as the development set to optimize the duration models, and the DEV00 set was used to test their performance. For the EVAL96 set we used the acoustic models from our 1996 Hub5 system, and for the DEV00 set we used the acoustic models from our 2000 Hub5 development system.

The duration models were used in the following way. For each utterance in the test set, we first generated a list of N−best hypotheses. We performed forced alignment of the utterance against each hypothesis in the N−best list. We scored the hypotheses using the alignments and the word duration models. The duration score was weighted and added to the acoustic and language model scores, and the hypothesis with the highest combined score was chosen.

With a view to optimize the performance of the word duration models, we conducted five experiments examining (1) the right back off models, (2) pause context modeling, and (3) ROS normalization.

3.1 Unnormalized durations

We examined the improvements in word error rate (WER) due to word duration models. For this, we trained the duration models using unnormalized durations. Triphone/phone back off was used to handle unseen words, and models were trained for pause context.

Model	Test Set	
	EVAL96	DEV00
Baseline	57.60%	42.80%
+ Duration Models	56.80%	42.40%

The results for both test sets show an improvement in the WER from word duration modeling.

3.2 Comparison of back off schemes

We examined two back off schemes: (1) a triphone back off followed by a phone back off for unseen triphones and (2) only a phone back off.

Model	Test Set	
	EVAL96	DEV00
Baseline	57.60%	42.80%
+ Duration Models (word/triphone/phone)	56.80%	42.40%
+ Duration Models (word/phone)	57.00%	42.50%

The results showed that a word/triphone/phone back off was better than a simple word/phone back off. Considering that we observed the triphone back off for only 4% of the words and phone back off for 1%, this result shows that the triphone durations carry significant information.

3.3 Modeling pause context

We compared the performance of modeling words in pause context with no pause context modeling. The aim of the pause context modeling was to capture the duration variations in a word due to 'prepausal lengthening'.

Model	Test Set	
	EVAL96	DEV00
Baseline	57.60%	42.80%
+ Duration Models (with pause context)	56.80%	42.40%
+ Duration Models (without pause context)	56.90%	42.50%

The results show a small improvement due to modeling pause context.

3.4 Normalized durations

The previous experiments examined the improvements in WER due to word duration modeling. However, in all of them, we modeled the raw duration. We know that different speakers speak at different rates, and the rate variations affect sounds differently. For example, when a person speaks at a fast rate, the vowels are reduced in duration much more than consonants. In view of this, we tried to normalize the durations of phones for ROS variations.

To perform ROS normalization, we estimated the average phone duration and normalized the individual phone durations using the average. The ROS normalization was done in training and testing. We also estimated the average phone duration at both the hypothesis and speaker levels and compared their performance.

Model	Test Set	
	EVAL96	DEV00
Baseline	57.60%	42.80%
+ Duration Models (unnormalized)	56.80%	42.40%
+ Duration Models (hypothesis level norm.)	56.60%	42.70%
+ Duration Models (speaker level norm.)	56.60%	42.20%

The results showed that normalized duration models perform better than unnormalized duration models. It was also observed that hypothesis−level ROS normalization was not consistent across different test sets whereas speaker−level ROS normalization resulted in consistent improvement for both test sets.

3.5 Class Based Normalization

We also experimented with using different ROS normalizations for different classes of sounds. The following table shows the results of our experiments for (1) global, (2) 3−class (pause, vowel, consonant) and (3) 7−class (pause, noise, vowel, stop, fricative, nasal, retroflex) ROS normalizations.

Model	Test Set	
	EVAL96	DEV00
Baseline	57.60%	42.80%
+ Duration Models (global ROS)	56.60%	42.20%
+ Duration Models (3−class ROS)	56.50%	42.00%
+ Duration Models (7−class ROS)	56.60%	42.00%

The results showed that class ROS normalization improves the performance of the duration models.

3.6 Improvements from larger N−best lists

We examined the improvements obtained from the use of duration models with larger N−best lists. The following table shows the WERs for varying N−best lists. For this experiment we used only the male subset of the DEV00 set.

Model	WER (DEV00 male subset)
Baseline	38.80%
+ Duration models (N=100)	37.90%
+ Duration models (N=200)	38.00%
+ Duration models (N=500)	37.70%
+ Duration models (N=2000)	37.50%

The results showed that the gains from duration models increase with increasingly larger N−best lists.

The duration models were used in SRI's 2000 Hub5 evaluation system [7] for rescoring the N−best lists. They resulted in a 0.6% improvement in WER.

4. SUMMARY

We proposed a new representation for word−level duration and used it to develop models for word duration. We also examined various issues related to these models and proposed solutions for them. We then performed experiments to study the reductions in the WER of a speech recognition system using the word duration models. The results showed that the word duration models produced significant reductions in WER. The experiments also showed that it was important to perform normalizations for ROS variations across speakers and sound classes.

REFERENCES

1. Veilleux N.M., and Ostendorf M., Probabilistic Parse Scoring with Prosodic Information, Proc. ICASSP'93, pages II−51−54.
2. Hunt A.J., A Generalised Model for Utilising Prosodic Information in Continuous Speech Recognition, Proc. ICASSP'94, pages 169−172.
3. Hieronymus J.L., McKelvie D., and, McInnes F.R., Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition, Proc. ICASSP'92, pages I−225−227.
4. Stolcke A., Shriberg E., Hakkani−Tür D., and Tür G., Modeling the Prosody of Hidden Events for Improved Word Recognition, Proc. 6th European Conference on Speech Communication and Technology, Budapest, Hungary, 1999.
5. Chung G. and Seneff S., A Hierarchical Duration Model for Speech Recognition Based on the ANGIE Framework, Speech Communication 27 (1999), 113−134.
6. Venkata Ramana Rao Gadde, Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani−Tür and Gokhan Tür, Prosody Modeling for Speech Recognition and Understanding, Proc. Hub5 Conversational Speech Understanding Workshop, Baltimore, 1999.
7. Stolcke A., Bratt H., Butzberger J., Franco H., Venkata R.G., Plauche M., Richey C., Shriberg E., Sonmez K., Weng F., and Zheng J., The SRI March 2000 Hub−5 Conversational Speech Transcription System, Proc. Speech Transcription Workshop, Univ. of Maryland, May 2000.